**تمرین عملی دوم تحلیل رگرسیون**

**اساتید:**

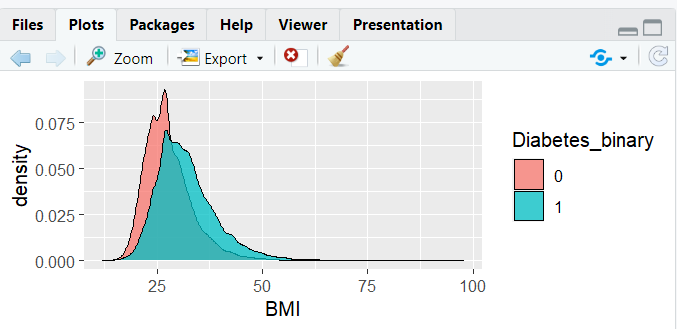
**دکتر میرصادقی**

**مهندس کدیور**

سوال ۱)

بله، با کمک فیچر‌ها می‌توان تقریب خوبی از اینکه فرد مبتلا به دیابت هست یا نه پیدا پیدا کرد که از همه‌‌ی این فیچرها برخی، تاثیرگذار‌تر هستند از جمله BMI و HeartDiseaseorAttack و HighBP.

اما بعضی از آن‌ها از جمله Sex تاثیرگذار‌ی کمتری دارند. در واقع انگار جنسیت رو میزان ابتلا به دیابت تاثیر ندارد.



A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

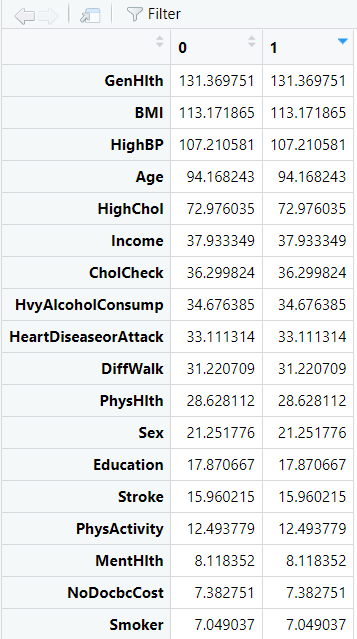
A screenshot of a computer

Description automatically generated with medium confidence

سوال ۲)

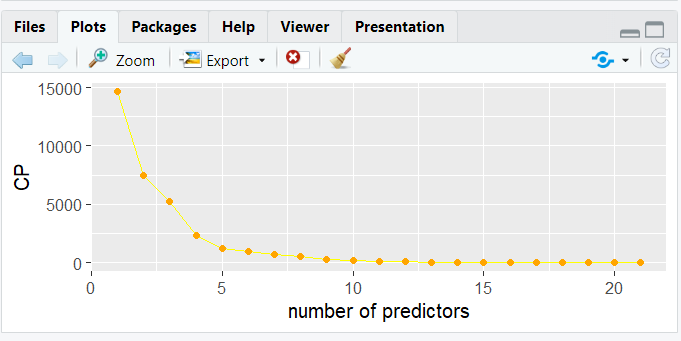
به کمک رندوم فارست فیچرهایی که تاثیرگذاری بیشتری دارند رو انتخاب می‌کنیم. در نتیجه آن‌ها بیشترین قدرت پیشبینی را نیز دارند.

با توجه به ارقام بدست آمده از GenHlth تا HighChol تاثیرگذارترین‌ها هستند.



سوال ۳)

با استفاده از subset selection زیر مجموعه‌ای از فیچر‌ها را انتخاب می‌کنیم. در واقع به کمک cp که یک روش فیچر سلکشن می‌باشد در می‌یابیم که حدودا ۱۱ فیچر تاثیرگذار داریم. حال به کمک best subset selection آن ۱۱ تای برتر رو خروجی میگیریم.



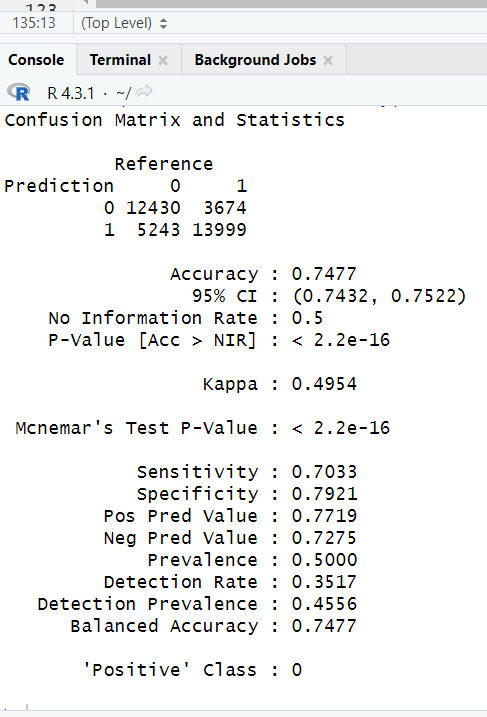
A screenshot of a computer

Description automatically generated with medium confidence

سوال ۴)

از آنجا که تعداد داده‌ها زیاد است و تعداد فیچرها نیز بالاست رندوم فارست یکی از بهترین گزینه‌ها می‌باشد چرا که split رندوم فارست به گونه‌ای است که آن‌هایی که correlation دارند با هم در یک درخت انتخاب نمی‌شود.

در این صورت sensitivity = 0.70و specificity = 0.79 می‌باشد



سوال ۵)

برای پیشبینی با دقت قابل قبول و با در نظر گرفتن توان محاسباتی می‌توان زیرمجموعه‌ای از best subset selection با بیشترین تاثیر را انتخاب کرد بطوری که مرورگر بتواند آن تعداد سوال رو هندل کند و روی آن random forest می‌زنیم و به کمک آن پیشبینی می‌کنیم.(البته با توجه به بخش دو میتوان گفت که حداقل 5 فیچر باید انتخاب شود.)